



BUSINESS DEVELOPMENT BY USING PREDICTION MODEL

By
**Md. Mifat
Rahman**
ID: 161-35-1526

This Report Presented in Fulfillment of the Requirements for the Degree of B.Sc.
in Software Engineering

Supervised By

Md. Shohel Arman
Lecturer

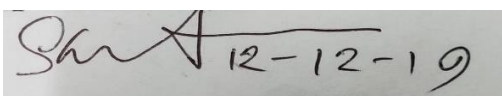
DEPARTMENT of SOFTWARE ENGINEERING
DAFFODIL INTERNATIONAL UNIVERSITY

FALL 2019

THESIS DECLARATION

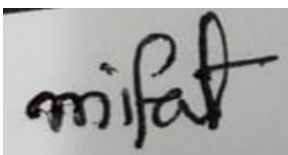
I hereby declare that, this thesis report is done by me under the supervision of Md. Shohel Arman, Lecturer, Department of Software Engineering, Daffodil International University, in partial fulfillment my original work. I am also declaring that neither this thesis nor any part therefore has been submitted else here for the award of Bachelor or any degree.

Supervised by



Md. Shohel Arman
Lecturer
Department of Software Engineering
Daffodil International University

Submitted by



Md. Mifat
Rahman
ID: 161-35-1526
Department of Software Engineering
Daffodil International University

ACKNOWLEDGEMENT

I am grateful to my creator for giving me the opportunity to complete this research work and learn so much. I am thankful to my research supervisor, Md. Shohel Arman, for providing careful guidance starting from selecting the research scope to successfully finalizing the research work. I would also like to thank Md. Sanzidul Islam, Lecturer, SWE for her valuable comments which was always insightful. Finally, I want to express my gratitude to Professor Dr. Touhid Bhuiyan, Head of the Software Engineering faculty, for inspiring us in all means. I am also thankful to all the lecturers, Department of Software Engineering who sincerely guided me at my difficulty. I am grateful to my parents for their unconditional support and encouragement. I am thankful to my friends who supported me throughout this venture.

Table of Contents

| | |
|---|------|
| THESIS DECLARATION | i |
| ACKNOWLEDGEMENT | ii |
| Table of Contents | iii |
| LIST OF TABLES | vi |
| LIST OF FIGURES..... | vii |
| LIST OF ABBREVIATIONS..... | viii |
| ABSTRACT | ix |
| CHAPTER 1 | 1 |
| INTRODUCTION | 1 |
| 1.1 Background | 1 |
| 1.2 Motivation of the Research | 2 |
| 1.3 Problem Statement..... | 2 |
| 1.4 Research Questions..... | 2 |
| 1.5 Research Objective | 3 |
| 1.6 Research Scope..... | 3 |
| 1.7 Thesis Organization | 3 |
| CHAPTER 2 | 4 |
| LITERATURE REVIEW | 4 |
| 2.1 Previous literature..... | 4 |
| 2.2 Previous research on algorithm performance on Business invest Data..... | 4 |
| 2.3 Previous research on invest growth Forecasting | 5 |
| 2.4 Research Gap..... | 5 |
| 2.5 Summary | 7 |
| CHAPTER 3 | 8 |
| RESEARCH METHODOLOGY..... | 8 |
| 3.1 Framing the Question..... | 8 |

| | |
|--|----|
| 3.2 Sample Size..... | 8 |
| 3.3 Data Collection | 8 |
| 3.4 Performing Analysis | 9 |
| 3.5 Model Development and Visualization | 9 |
| 3.6 Model Evaluation Metrics | 9 |
| 3.7 In-Sample Evaluation | 10 |
| 3.8 Out-of-Sample Evaluation | 10 |
| 3.9 Summary..... | 10 |
| CHAPTER 4..... | 11 |
| RESULTS AND DISCUSSIONS | 11 |
| 4.1 Data Analysis Technique | 11 |
| 4.1.1 Simple Linear Regression..... | 11 |
| 4.1.2 Multiple Linear Regression..... | 11 |
| 4.1.3 Polynomial Regression | 12 |
| 4.2 Evaluation Metrics | 13 |
| 4.2.1 R-Squared | 13 |
| 4.2.2 Mean Squared Error | 13 |
| 4.3 Correlation Visualization..... | 14 |
| 4.3.1 Yearly Investment | 14 |
| 4.3.2 Yearly Profit | 14 |
| 4.4 Distribution Plots (In-Sample) | 15 |
| 4.4.1 SLR..... | 15 |
| 4.4.2 MLR..... | 15 |
| 4.4.3 Polynomial fit..... | 16 |
| 4.5 Evaluation Metrics Table (In-Sample) | 17 |
| 4.6 Conclusion from In-Sample Evaluation | 17 |
| 4.7 Distribution Plots (Out-of-Sample)..... | 18 |
| 4.7.1 SLR (Training Set) | 18 |

| | |
|--|----|
| 4.7.2 SLR (Testing Set)..... | 18 |
| 4.7.3 MLR (Training Set)..... | 19 |
| 4.7.4 MLR (Testing Set)..... | 19 |
| 4.7.5 Polynomial fit (Training Set)..... | 20 |
| 4.7.6 Polynomial fit (Testing Set)..... | 20 |
| 4.8 Evaluation Metrics Table (Out-of-Sample) | 21 |
| 4.9 Conclusion from Out-Of-Sample Evaluation | 21 |
| CHAPTER 5 | 22 |
| CONCLUSIONS AND RECOMMENDATIONS | 22 |
| 5.1 Findings and Contributions..... | 22 |
| 5.2 Limitations | 23 |
| 5.3 Recommendations for Future Works..... | 23 |
| REFERENCES | 24 |

LIST OF TABLES

| | |
|---|----|
| Table 1: Literature Review | 5 |
| Table 2 In-Sample Evaluation..... | 17 |
| Table 3: Out-of-Sample Evaluation | 21 |

LIST OF FIGURES

| | |
|--|----|
| Figure 1: Valid investment & profit | 15 |
| Figure 2 : Valid profit and Year..... | 16 |
| Figure 3 : SLR In-Sample Distribution Plot..... | 17 |
| Figure 4 : MLR In-Sample Distribution Plot..... | 18 |
| Figure 5 : Polynomial In-Sample Distribution Plot | 19 |
| Figure 6 : SLR Out-of-Sample Distribution Plot (Training Set) | 21 |
| Figure 7 : SLR Out-of-Sample Distribution Plot (Testing Set) | 22 |
| Figure 8 : MLR Out-of-Sample Distribution Plot (Training Set)..... | 23 |
| Figure 9 : MLR Out-of-Sample Distribution Plot (Testing Set)..... | 24 |
| Figure 10 : Polynomial Out-of-Sample Distribution Plot (Training Set)..... | 25 |
| Figure 11 : Polynomial Out-of-Sample Distribution Plot (Testing Set)..... | 26 |

LIST OF ABBREVIATIONS

| Abbreviation | Explanation |
|---------------------|----------------------------|
| SLR | Simple Linear Regression |
| MLR | Multiple Linear Regression |
| MSE | Mean Squared Error |

ABSTRACT

In this research, we investigated business development by using the performance of regression algorithms namely Simple Linear Regression, Multiple Linear Regression and Polynomial Regression for predicting the number of Investment growth in business. We utilized the information of two years of speculation of foundation to anticipate the quantity of legitimate interest in every voting public. Foreseeing pace of venture from the aggregate sum of speculation is extremely vital for foundation Data Analytics. From past venture of foundation and year esteems, we can foresee speculation development of benefit throw and from those qualities, we can without much of a stretch infer speculation turnout, which is important for establishment partners. A high or low contribute turnout means that how much development benefit and stream a sum in an establishment. We used excel to carry out the all the research work such as data preprocessing, feature selection, data analysis and evaluation. We found that Multiple Linear Regression is the best performer among Simple Linear Regression and Polynomial Regression both in terms of In-Sample and Out-of-Sample evaluation. We reached to the conclusion that multiple attributes can contribute to less error prone prediction.

CHAPTER 1

INTRODUCTION

1.1 Background

Data Analytics in business plays a big role in every institution. The first forecast to use big data analytics was in 1990, at the US business. A total of \$7 billion was spent on the many business strategies and about 55% of this amount was wasted of unexpected investment. Using Big Data, analysis can measure possible business profit. In business Previous Investment, profit analysis and predict possible upcoming year how will be the flow of investment & profit. With data mining, Big Data, machine learning traditional models plays a significant role for prediction. Data mining for institutions helps to know the investment, profit & ratio and what was the flow of curve .Different algorithm is using in data mining, so we can easily predict and bring out which algorithm is best.

1.2 Motivation of the Research

Business information investigation assumes a significant job in understanding the general monetary circumstance, conduct of benefits and which contribute make an ideal proportion. This kind of analysis helps institution to assess stakeholder's interest and trust in business. Representative gatherings can utilize this data to mastermind great contribute crusades to bring issues to light and energy among individuals which can prompt a superior social and business circumstance everywhere throughout the nation in any business.

1.3 Problem Statement

We found that information investigation of beginning Business gauge has not been done in contrast with different nations. We found a dataset of USA speculation development figure on Kaggle. We looked if any investigation was done yet didn't discover any. We knew the types of researches done with invest data of other countries and so we decided to do a similar research which can help the investment of business and people in the way that other researches have already done.

1.4 Research Questions

The research question was

- RQ1: Which regression algorithm performs best in predicting number of investment?
- RQ2: Is there only one or more than one variables that positively influence the prediction?

1.5 Research Objective

The objectives of this research are

- To discover the best performing regression algorithm
- To explore whether one variable or more than one variable is needed for predicting the outcome.
- To discover best accuracy.
- To predict better result.

1.6 Research Scope

We identified the research gap in “Rasheed, A. (2016). Data Mining Application into Business Investment growth in USA Business with Regression Analysis. Journal of Asian Scientific Research, 2(16), 893.” where we saw that lone Simple Linear Regression was utilized and just one credit was talked about going to have contributed in anticipating the quantity of members. So in the wake of getting a comparative dataset of USA business we picked three relapse calculations to anticipate substantial number of benefit from the accessible dataset lastly assess them utilizing two famous assessment measurements of relapse calculations.

1.7 Thesis Organization

In the following, in second chapter we examined about different looks researches done on the same subject including the research gap. In chapter three, we examined our proposed research methodology. In chapter four, we discussed the analysis results. Finally in chapter five, we discussed the conclusions and recommendations containing findings, limitations and future work.

CHAPTER 2

LITERATURE REVIEW

2.1 Previous literature

A study revealed that there is a significant relationship between year and number of invest (Ojaganju, Hukaila et al, 2012). They used linear regression model to establish a relation between the two variables. Another study done in anticipating the political behavior of candidates revealed that KNN performs with higher accuracy than Naïve Bayes and Decision tree (Amun Dabazadeh et al, 2013). Also in another study in Australian Electoral Commission done in 2016 revealed that there is a clear relation between profit & investment. Another study was done to forecast investment growth results using two layer perceptron (G.S Gill, 2005). Another study was done to study business trend in the US which revealed that there is clear difference between how profit & invest (Sikha Magui, 2007).

2.2 Previous research on algorithm performance on Investment growth

Algorithm performance evaluation research has been done to anticipate political behavior (Amun Babazadeh Sangar et al, 2013). In this examination they utilized Naïve Bayes calculation, Decision Tree and KNN calculation to foresee voter investment in political decision. They used data of 100 qualified persons eligible to participate in the election and concluded that KNN in comparison to Naïve Bayes and Decision tree, performs better. They used the tool Orange and the CRISP data mining method to carry out their research.

2.3 Previous research on Invest Growth Forecasting

A two layer neural system was utilized to estimate contribute development results (G.S. Gill 2005). They utilized master reactions to a battery of ten inquiries and it was utilized to prepare the neural system. A two layer arrange including info and yield layers was utilized. The general business from 1977 to 2004 was utilized to prepare the model and the 2004 outcome was utilized for testing. They inferred that the information was not adequate and that thinks about yield. The researchers believe that more expert responses are needed to make more accurate predictions but the responses should not be idiosyncratic otherwise the network will also reflect those traits.

2.4 Research Gap

In the following table 1 we provided the information of research gap and based on this gap we continued our research.

Table 1: Literature Review

| Paper | Year | Author | Objective | Data | Methodology | Conclusion |
|---|------|--|---|--|-------------------|---|
| Data Mining Application in Business Investment Trends in USA with regression analysis | 2012 | Ojaganju, Hukaila, Tomori, Adekola Rasheed | Providing a basic model which relates invest of business USA Market's with periods of forecast. | Raw-Data for business USA Invest growth 1932 to 2010 | Linear regression | There is a significant relationship between profit and the number of investors. |

| Paper | Year | Author | Objective | Data | Methodology | Conclusion |
|---|------|--|---|---|---|---|
| Participation anticipation in invest using data mining | 2013 | Amun Babazadeh Sangar, Syyed Reza Khaze, Laya Ebrahimi | Anticipating the investor behavior of stakeholders | Data of 100 qualified institutions | KNN, Classification Tree, Naïve Bayes | KNN performs with higher accuracy than the other two algorithms |
| 2016 Business of representatives and giant market | 2016 | Australian Ecommarce | analyses of business investment 2016 held | 2016 Business investors | | 1. There is a clear relationship between profit and investors turnout at the national level |
| Data Mining techniques to study business trend patterns in the US | 2007 | Sikha Bagui, Dustin Mink, Patrick Cash | Study business patterns in the United States business Representatives and shows how the results can be interpreted. | the raw data available at http://clerk.house.gov | t-weight calculations, association rule mining, decision tree | From the t-weight conclusions we can see that, except for Issue 585, there is a difference in how Profit, income, loss Invest . Also same for association rule mining, decision tree. |

| | | | | | | |
|---|------|----------------------------------|---|--|----------------------|---|
| Invest Growth Forecasting using two layer perceptron | 2005 | G.S Gill | Using ANN to forecast business investment | All general businesses between 1996-2004 | Two layer perceptron | A generalized conclusion cannot be made because of insufficient data. |
| How to classify a Business: Can a neural network do it | 2005 | Antonio Caleiro | Classifying a Business using ANN | Business data | Perceptron | Perceptron can classify a Business provided that it is given sufficient data |
| Strategic profit in Proportional Representation Systems | 2007 | Ignacio Lago | Impact of district magnitude on strategic profit | 1990 and 1995 Spanish business market data | | Strategic business does not depend on rational expectations but on heuristics. |
| Trial invest forecasts of global market | 1990 | James E Campbell, Kenneth A Wink | This paper examines the trial invest forecasts of the two business market | | | The trial invest equation suggested by the most accurate in forecasting the Global market |

2.5 Summary

Our Study is motivated by the need for research focusing valid profit prediction using the available dataset. We are going to use three regression algorithms to build our models and predict and finally evaluate them and come up the best performing algorithm for this purpose and also see which variables are most effecting in predicting the outcome.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Framing the Question

In any data science research, we need to ask as many relevant questions as possible. We have to take a gander at the accessible dataset and make important suppositions and detail inquiries from it. We have to take a gander at the accessible dataset and make important suppositions and detail inquiries from it. We get measure of contribute and date. So we chose to target substantial number of contribute forecast from the accessible information. Our questions mainly were: which regression algorithm is best for predicting valid number of invests and is only one or more than one variable needed to get a more precise and accurate prediction of invest.

3.2 Sample Size

Our dataset contains information of two years of Invest in an establishment in USA. The three years are 2016 and 2017. For every year we have information of body electorate, so in absolute we had 200 columns of information. The information was at that point clean and in great structure.

3.3 Data Collection

The information was gathered from Kaggle. Various information researchers were likewise utilizing this information to show invest growth of business. The link of this data source is: <https://www.kaggle.com/saqibmujtaba/investment-growth-forecast>

3.4 Performing Analysis

We are going to see the correlation between the variables “valid number of votes” with the variable: “registered number of voters” and the variable “year”. At that point we are going to utilize relapse examination on the factors accessible on our dataset. The chosen algorithms are Simple Linear Regression, Multiple Linear Regression and Polynomial Regression. We are going to visualize the results with actual values and fitted values and compare them visually. Finally we are going to use R-squared and Mean Squared Error to evaluate results of the algorithms and conclude our findings.

3.5 Model Development and Visualization

We are going develop models for simple linear regression, multiple linear regression and polynomial regression. After developing the models, we are going to visualize the actual values with the fitted values. Visualization will help us evaluate the results.

3.6 Model Evaluation Metrics

We are going assess our models dependent on R-Squared worth and Mean Squared Error. We are going assess our models dependent on R-Squared worth and Mean Squared Error.

3.7 In-Sample Evaluation

Firstly we are going to evaluate the three models using in-sample evaluation, meaning that, we are going to train and test our model using the same data. So here we are essentially trying to see how well our models perform with the information that they have just been prepared with. This typically brings about a superior forecast or result as the models predict dependent on known information. However, all things considered, in situations, we ought to consistently test our model utilizing obscure information.

3.8 Out-of-Sample Evaluation

This is our second evaluation approach. In out-of-test assessment, we train our models with a specific measure of information and afterward test those models utilizing obscure information from the dataset. This is a superior methodology for assessing the models as this reveals to us how well the models perform on obscure cases. In out-of-test assessment, we train our models with a specific measure of information and afterward test those models utilizing obscure information from the dataset. This is a superior methodology for assessing the models as this discloses to us how well the models perform on obscure cases.

3.9 Summary

We are going to discover the best relapse calculation that predicts the benefits of venture, which is a vital aspect for finding an organization's circumstance. We are also going to find out the number of variables needed for making such predictions using data mining techniques. We are additionally going to discover the quantity of factors required for making such forecasts utilizing information mining systems.

CHAPTER 4

RESULTS AND DISCUSSIONS

4.1 Data Analysis Technique

Data analysis was done using python and machine learning framework scikit learn and Jupyter Notebook and Microsoft excel. Maximum Result got by using excel. Three regression algorithms were used for this research: Simple Linear Regression, Multiple Linear Regression and Polynomial Regression.

4.1.1 Simple Linear Regression

Simple Linear Regression establishes the relationship between two variables using a straight line. Straight relapse endeavors to draw a line that comes nearest to the information by finding the incline and catch that characterize the line and limit relapse mistakes.

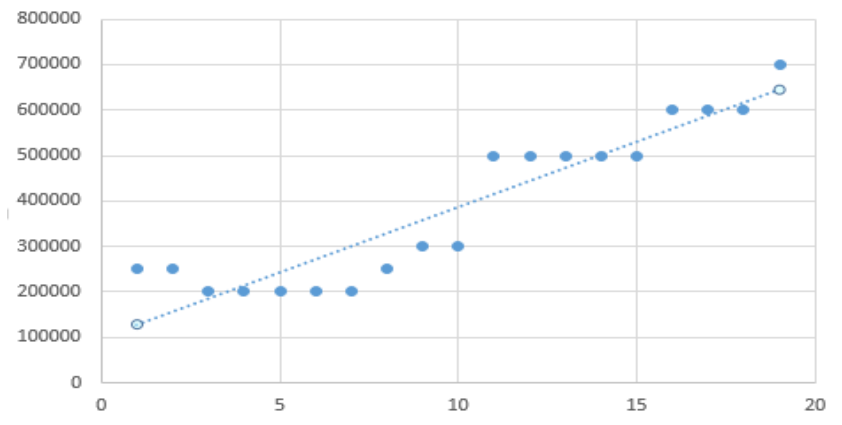


Figure 01: Simple Linear Regression

4.1.2 Multiple Linear Regression

In the event that at least two informative factors have a direct association with the reliant variable,

the relapse is known as a various straight relapse.

4.1.3 Polynomial Regression

Numerous information connections don't pursue a straight line, so analysts use non-direct relapse. The two are comparable in that both track a specific reaction from a lot of factors graphically. Yet, nonlinear models are more convoluted than direct models in light of the fact that the capacity is made through a progression of presumptions that may come from experimentation.

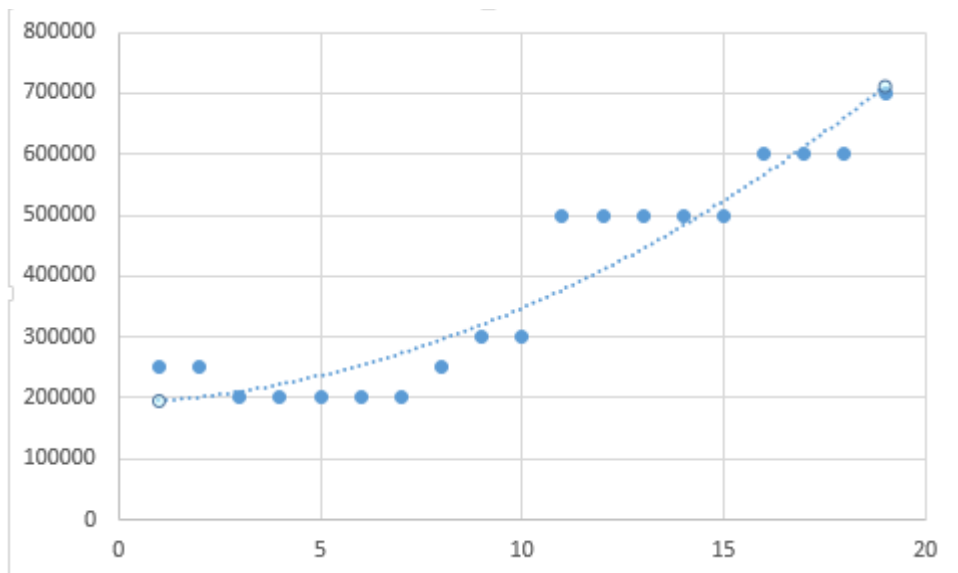


Figure 02: Polynomial Regression

4.2 Evaluation Metrics

We think about the genuine qualities and anticipated qualities to compute the exactness of a relapse model. Assessment measurements give a key job in the advancement of a model, as it gives knowledge to zones that require improvement.

4.2.1 R-Squared

R-squared is not error, but is a popular metric for accuracy of your model. It represents how close the data are to the fitted regression line. The higher the R-squared, the better the model fits your data. Best possible score is 1.0 and it can be negative (because the model can be arbitrarily worse).

$$R \text{ squared} = 1 - (RSS/TSS)$$

Where:

RSS is Residual Sum of Squares and

TSS is Total Sum of Squares.

4.2.2 Mean Squared Error

Mean Squared Error (MSE) is the mean of the squared mistake. It's more prevalent than Mean total blunder on the grounds that the center is equipped more towards enormous mistakes. This is because of the squared term exponentially expanding bigger blunders in contrast with littler ones.

Steps to calculate MSE:

Find the regression line.

- Insert your X values into the linear regression equation to find the new Y values (Y').

- ▶ Subtract the new Y value from the original to get the error.
- ▶ Square the errors.
- ▶ Add up the errors.
- ▶ Find the mean.

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y - \hat{f}(x_i))^2$$

4.3 Correlation Visualization

4.3.1 Investment & Profit

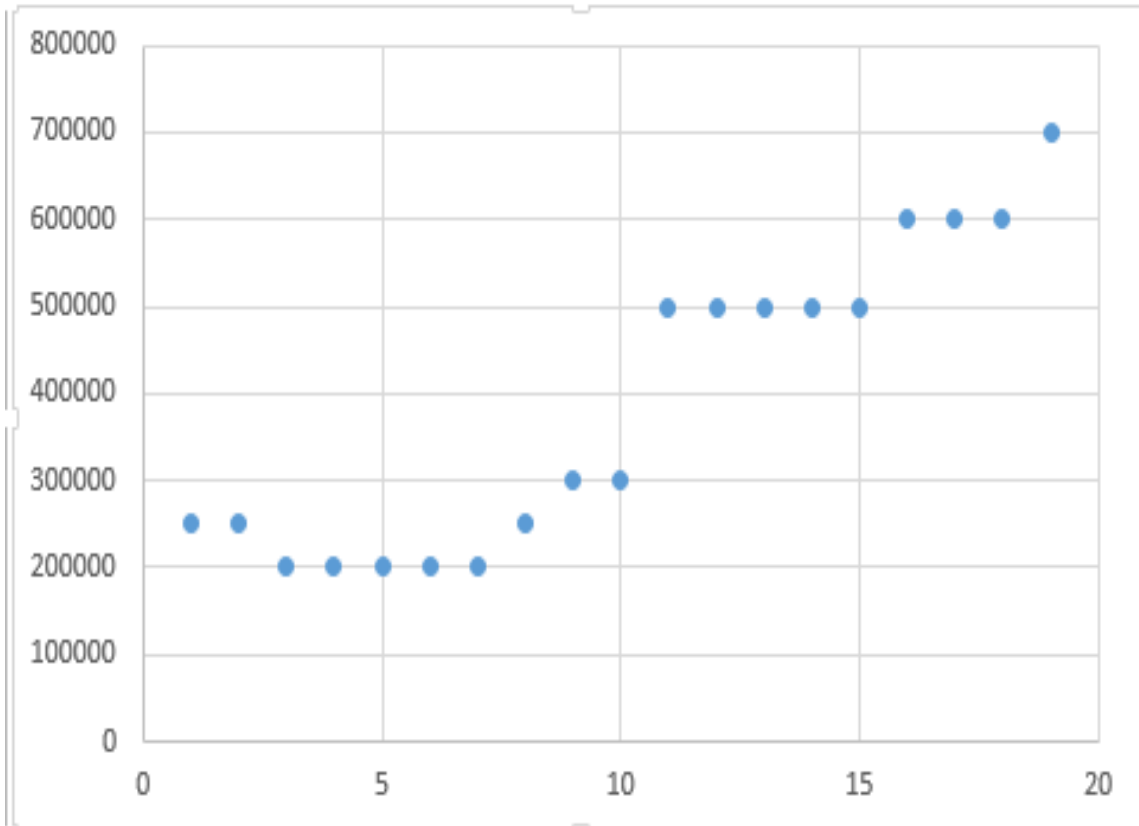


Figure 1: Investment and Profits

4.3.2 The two variables invest and profit has a positive linear relationship. So as relationship of date, profit, investment.

4.3.3 Valid Investment and Year

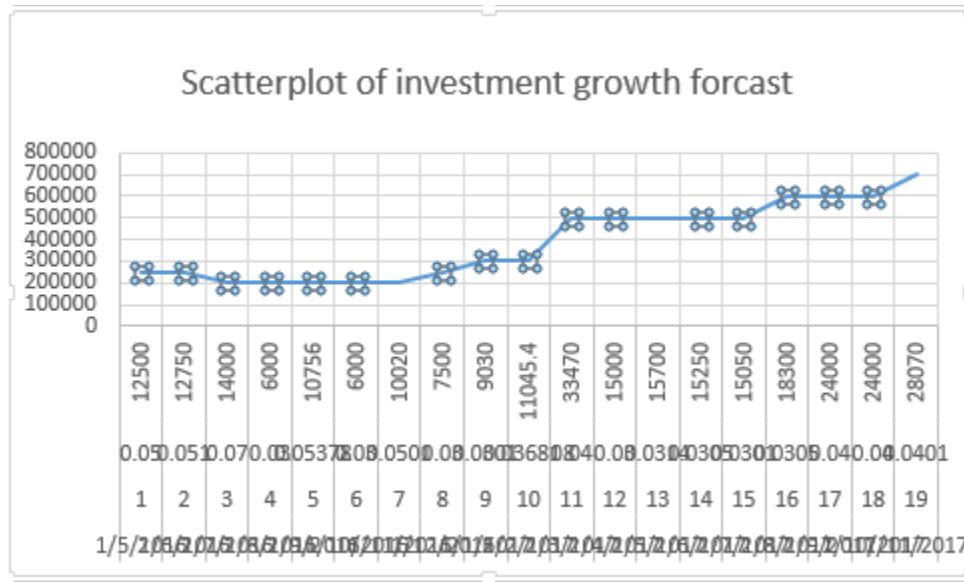


Figure 2: Investment vs Year growth

4.4 Distribution Plots (In-Sample)

4.4.1 SLR

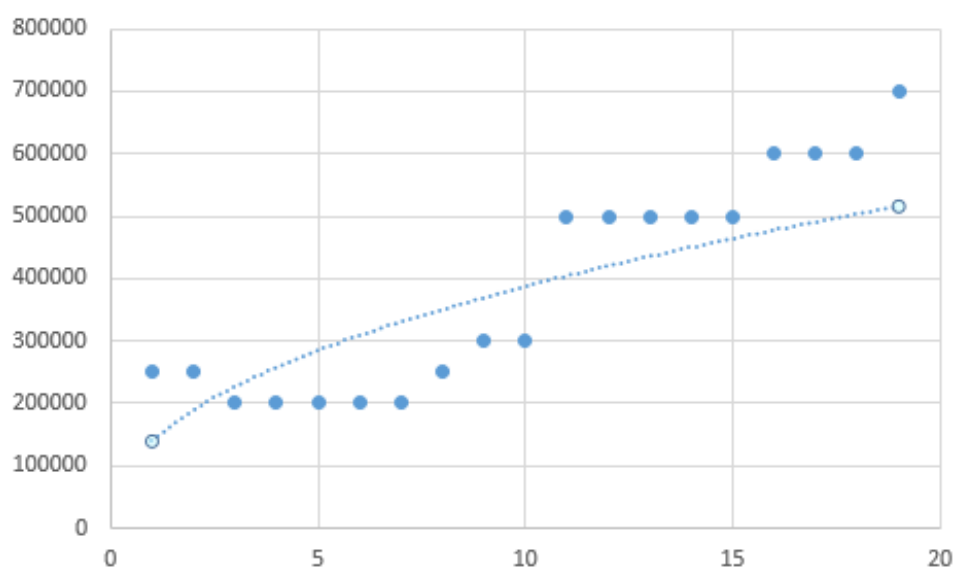


Figure 3 : SLR In-Sample Distribution Plot

SLR fit has a little bit of under-fitting issue here.

4.4.2 MLR

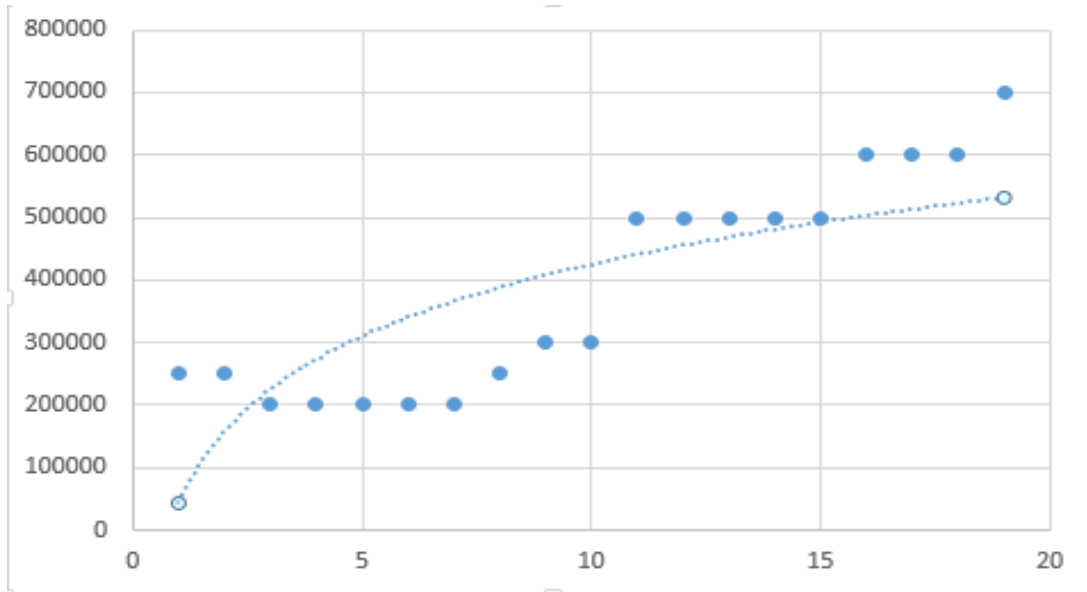


Figure 4 : MLR In-Sample Distribution Plot

MLR seems to be a good fit as a model.

4.4.3 Polynomial fit

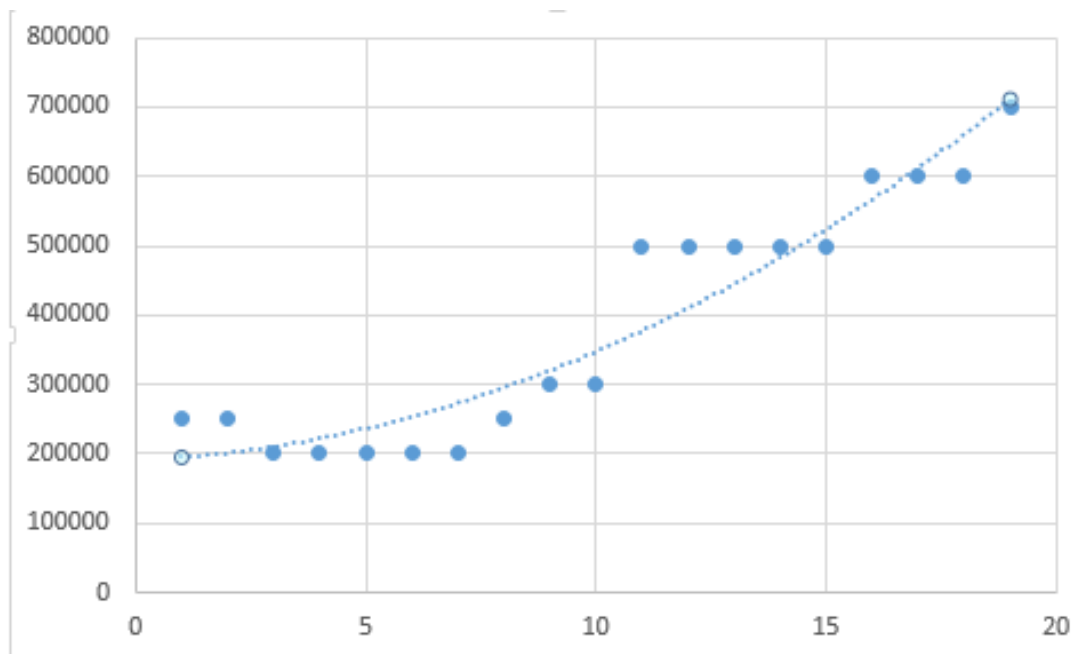


Figure 5 : Polynomial In-Sample Distribution Plot

Polynomial fit also seems to be a good fit as a model.

4.5 Evaluation Metrics Table (In-Sample)

Table 2 In-Sample Evaluation

| Model Name | R-Squared Value | MSE |
|-----------------------|------------------------|-------------------|
| SLR | 0.8474474644 | 601594362.1010092 |
| MLR | 0.8934647484 | 417323621.7448205 |
| Polynomial Fit | 0.8836454764 | 518392727.6093318 |

4.6 Conclusion from In-Sample Evaluation

In comparison, MLR has the highest R-Squared value and the lowest MSE. So MLR appears to be the best fit in In-Sample evaluation.

4.7 Distribution Plots (Out-of-Sample)

For Out-of-Sample evaluation, we had 200 test samples and 350 training samples.

4.7.1 SLR (Training Set)

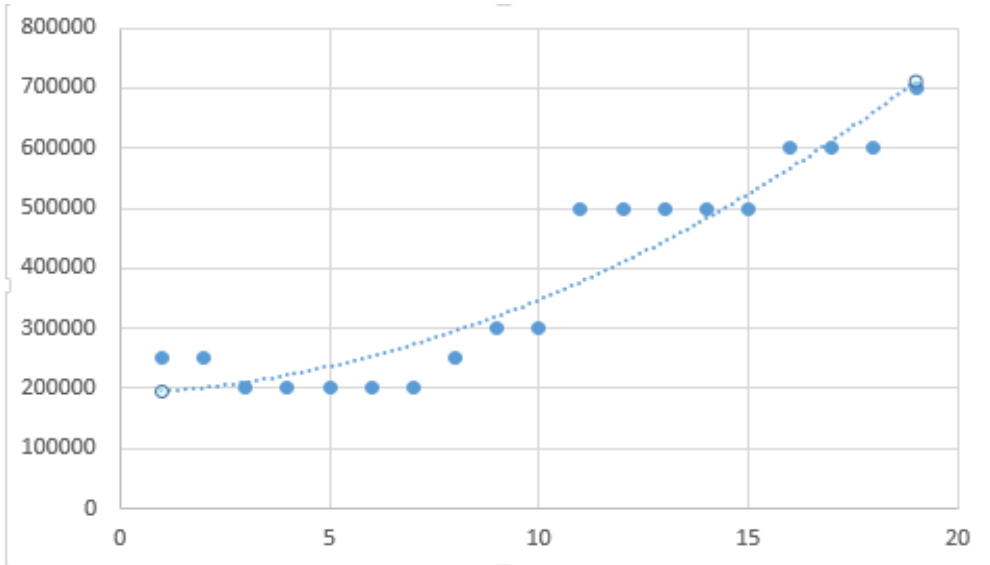


Figure 6 : SLR Out-of-Sample Distribution Plot (Training Set)

SLR seems to have a bit of under-fitting issue here.

4.7.2 SLR (Testing Set)

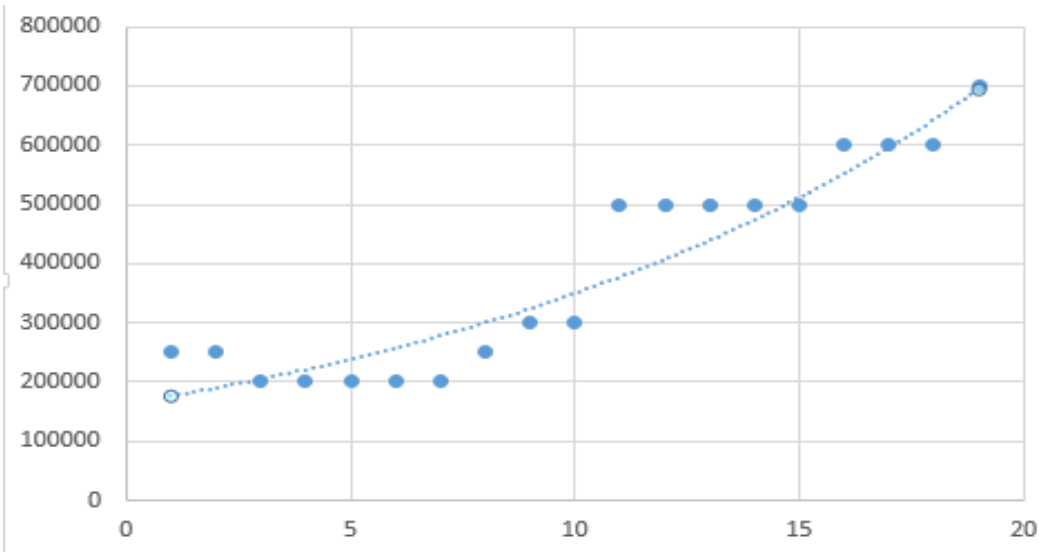


Figure 7 : SLR Out-of-Sample Distribution Plot (Testing Set)

Also, in the test data, it has under-fitting issue.

4.7.3 MLR (Training Set)

| Regression Statistics | | Multiple Linear Regression | | | | | | |
|-----------------------|-------------|----------------------------|----------|----------|----------------|-----------|-------------|-------------|
| Multiple R | 0.966886 | | | | | | | |
| R Square | 0.934869 | | | | | | | |
| Adjusted R Square | 0.916261 | | | | | | | |
| Standard Error | 50316.68 | | | | | | | |
| Observations | 19 | | | | | | | |
| ANOVA | | | | | | | | |
| | df | SS | MS | F | Significance F | | | |
| Regression | 4 | 5.09E+11 | 1.27E+11 | 50.23819 | 3.75E-08 | | | |
| Residual | 14 | 3.54E+10 | 2.53E+09 | | | | | |
| Total | 18 | 5.44E+11 | | | | | | |
| | Coefficient | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
| Intercept | -3269509 | 5872649 | -0.55673 | 0.586499 | -1.6E+07 | 9326071 | -1.6E+07 | 9326071 |
| Deposit | 79.65026 | 138.312 | 0.575874 | 0.573842 | -217 | 376.3 | -217 | 376.3 |
| Months | 17745.87 | 4625.084 | 3.836875 | 0.001814 | 7826.053 | 27665.69 | 7826.053 | 27665.69 |
| PFT_Perc | -1191763 | 1562284 | -0.76283 | 0.458224 | -4542529 | 2159003 | -4542529 | 2159003 |
| PFT | 8.790056 | 2.657947 | 3.307084 | 0.005189 | 3.089326 | 14.49079 | 3.089326 | 14.49079 |

Figure 8 : MLR Out-of-Sample Distribution Plot (Training Set)

MLR appears to have a good fit for the training data.

4.7.4 MLR (Testing Set)

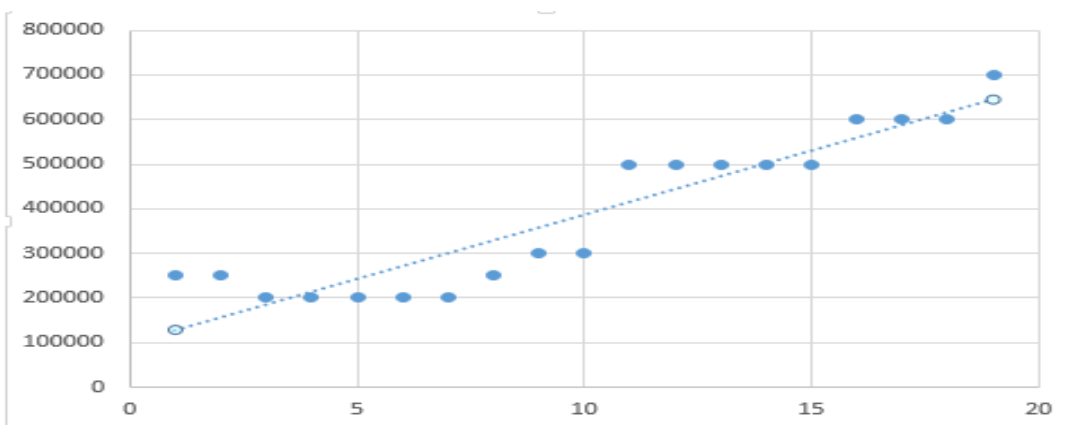


Figure 9 : MLR Out-of-Sample Distribution Plot (Testing Set)

MLR still has a good fit with the test data.

4.7.5 Polynomial fit (Training Set)

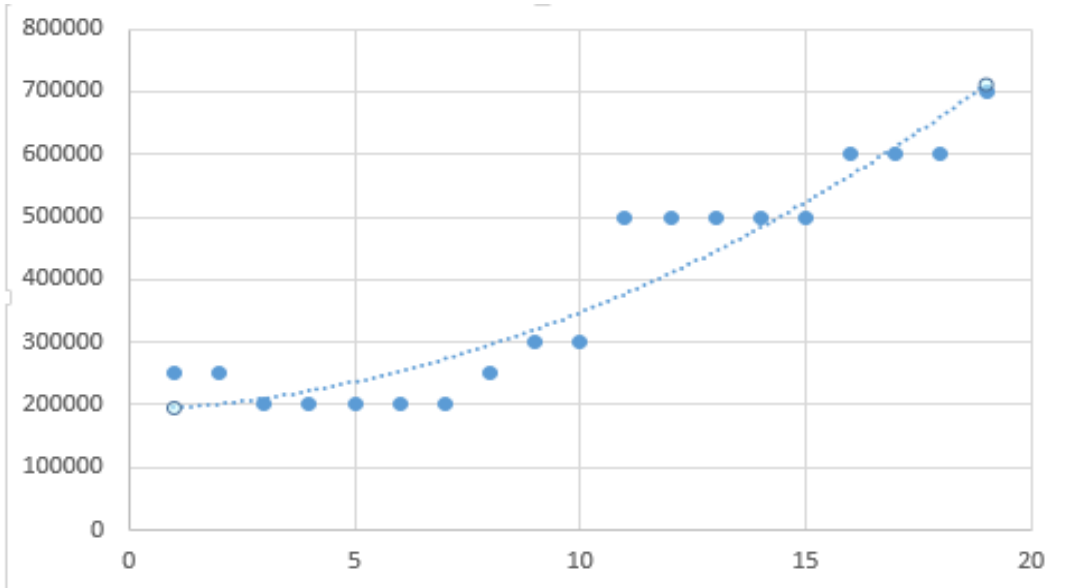


Figure 10 : Polynomial Out-of-Sample Distribution Plot (Training Set)

Polynomial fit seems to have a slightly under-fitting issue here.

4.7.6 Polynomial fit (Testing Set)

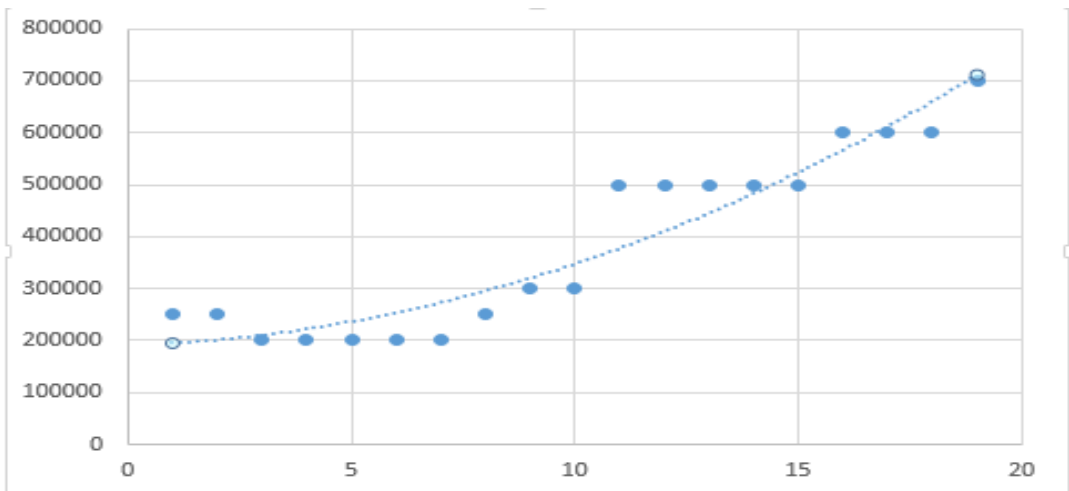


Figure 11 : Polynomial Out-of-Sample Distribution Plot (Testing Set)

Polynomial fit still has some slight under-fitting issue here on the test data.

4.8 Evaluation Metrics Table (Out-of-Sample)

Table 3: Out-of-Sample Evaluation

| Model Name | R-Squared Value | MSE |
|-----------------------|------------------------|--------------------|
| SLR | 0.8463870279543719 | 546563482.7320896 |
| MLR | 0.93486985485 | 383676328.92809874 |
| Polynomial Fit | 0.8575726115743108 | 506764555.2177598 |

4.9 Conclusion from Out-Of-Sample Evaluation

In comparison, MLR has the highest R-squared value among all three of the models. So, MLR is the best in out-of-sample evaluation.

CHAPTER 5

CONCLUSIONS AND RECOMMENDATIONS

5.1 Findings and Contributions

The research goal of this study was to investigate which algorithms performs best in predicting business invest growth. We had data of two years of institution in USA. For three hundred constituency each year, we had total 400 hundred rows of data. We tried to fit the data to out regression models and predict the outcome “valid number of investment” from the predictor variables “year” and “profit”. For In-Sample evaluation, we trained our model with 200rows of data and predicted the outcome and evaluated thereby. In the case of Out-of-Sample Evaluation, we had 350rows of data as training data and 200rows of test data to evaluate our trained models. For SLR and polynomial regression, the predictor variable was “invest” and for MLR the predictor variables were “profit” and “year”. We also visualized the correlation between “year” and “investment” and also between “profit” . In both cases, it showed that both the variables had strong positive linear relationship with the variable “invest” which was an early indicator that regression algorithm that takes into account both the variables is going to predict with good level of accuracy. Among SLR, MLR and Polynomial regression, considering from both In-Sample and Out-of-Sample evaluation, MLR outperformed SLR and Polynomial Regression. It stood out with its less error rate and R-squared value from both SLR and Polynomial Regression. As MLR was found to be the best which also tells us that more than one variables are needed to accurately predict valid votes.

5.2 Limitations

There are a few restrictions to be recognized. The quantity of factors were two for expectation; nonetheless, if there were different factors too, we could utilize them too and think of far superior research questions and better forecasts.

5.3 Recommendations for Future Works

Future research ought to be coordinated towards inspecting how different calculations perform on this dataset. A progressively thorough research on calculation execution should be possible on this dataset to concoct fascinating new outcomes which could assist better with understanding the extent of more work on political decision information investigation.

REFERENCES

1. Rasheed, A. (2012). Data Mining Application into Potential Voters Trends in Usa Elections with Regression Analysis. *Journal of Asian Scientific Research*, 2(12), 893.
2. Sangar, A. B., Khaze, S. R., & Ebrahimi, L. (2013). Participation anticipating in elections using data mining methods. arXiv preprint arXiv:1307.7429.
3. Bagui, S., Mink, D., & Cash, P. (2007). Data mining techniques to study voting patterns in the US. *Data Science Journal*, 6, 46-63.
4. Lago, I. (2008). Rational expectations or heuristics? Strategic voting in proportional representation systems. *Party Politics*, 14(1)
5. Prof. Anupama George and Prof. Rupashree R, *International Journal of Research in Engineering, IT and Social Sciences*, ISSN 2250-0588, Impact Factor: 6.565, Volume 09 Issue 04, April 2019, Page 22-26
(7) (PDF) Forecast of Foreign Direct Investment Inflow (2019-2023) with reference to Indian Economy. Available from:
https://www.researchgate.net/publication/333718739_Forecast_of_Foreign_Direct_Investment_Inflow_2019-2023_with_reference_to_Indian_Economy
6. Gill, G. S. (2008). Election result forecasting using two layer perceptron network. *Journal of Theoretical and Applied Information Technology*, 4(11), 1019-1024.
7. Traugott, M. W., & Tucker, C. (1984). Strategies for predicting whether a citizen will

vote and estimation of electoral outcomes. *Public Opinion Quarterly*, 48(1B), 330-343

8. Reif, K., & Schmitt, H. (1980). Nine second-order national elections—a conceptual framework for the analysis of European Election results. *European journal of political research*, 8(1), 3-44.

9. Tumasjan, A., Sprenger, T. O., Sandner, P. G., & Welpe, I. M. (2011). Election forecasts with Twitter: How 140

10. Caleiro, A. (2005). How to Classify a Government? Can a neural network do it? (No. 2005/09). Documento de Trabalho.

11. Larose, D. T., & Larose, C. D. (2014). *Discovering knowledge in data: an introduction to data mining*. John Wiley & Sons.

12. Lago, I. (2008). Rational expectations or heuristics? Strategic voting in proportional representation systems. *Party Politics*, 14(1)

13. Larose, D. T., & Larose, D. T. (2006). *Data mining methods and models* (Vol. 12). Hoboken (NJ): Wiley-Interscience.

14. Olagunju, M., 2009. Evaluation of students' performances in an examination using data mining techniques. 1: 116-126.

15. Campbell, J. E., & Wink, K. A. (1990). Trial-heat forecasts of the presidential vote. *American Politics Quarterly*, 18(3), 251-269.

16. Lavanya, D., & Rani, K. U. (2011). Performance evaluation of decision tree classifiers on medical datasets. *International Journal of Computer Applications*, 26(4), 1-4.

17. Dokas, P., Ertoz, L., Kumar, V., Lazarevic, A., Srivastava, J., & Tan, P. N. (2002, November). Data mining for network intrusion detection. In *Proc. NSF Workshop on Next Generation Data Mining* (pp. 21-